

BAB I

PENDAHULUAN

1.1 Latar Belakang

Teknologi yang berkembang pesat pada abad ke 21 telah mempermudah cara kerja manusia sebagai pengembang atau penemunya. Bahkan saat ini, manusia telah mengembangkan teknologi ke tahap dimana teknologi tersebut dapat belajar secara “mandiri” yaitu *machine learning*. *Machine learning* atau pembelajaran mesin adalah cabang pembelajaran yang berfokus belajar mengenai data secara “mandiri” tanpa harus diprogram terus menerus oleh manusia. Sehingga, *machine learning* membutuhkan data yang valid sebagai bahan belajar ketika berjalannya proses *training* sebelum akhirnya digunakan ketika *testing* untuk hasil yang optimal [1].

Salah satu cabang dari *machine learning* adalah *data mining* yang merupakan pemrosesan untuk mengolah data sehingga mampu mendapatkan informasi. Secara sederhana, *data mining* adalah sebuah langkah untuk bisa mengekstraksi data dan menjadikannya informasi yang bermanfaat namun bersifat implisit atau belum diketahui [2].

Data mining memiliki ragam teknik di dalamnya untuk bisa memproses sebuah data menjadi informasi, salah satunya yaitu teknik *clustering*. *Clustering* adalah klasifikasi dokumen tanpa pengawasan yang membagi koleksi teks menjadi beberapa subset yang disebut dengan kluster. *Clustering* sering dilakukan sebagai langkah awal dalam proses *data mining*, *clustering* juga memiliki banyak algoritma yang bervariasi seperti *K-Means*, *Improved K-Means*, *K-Medoids (PAM)*, *Fuzzy C-means*, *DBSCAN*, *CLARANS* dan *Fuzzy Subtractive* [3].

Algoritma *K-Means* adalah algoritma *partitioning clustering* yang populer digunakan dan dianggap memiliki kinerja yang baik. Cara kerja dari algoritma ini adalah dengan menentukan kelompok *centroid* (titik pusat) pertama yang dipilih secara acak sebagai titik awal kluster, kemudian dilakukan pengujian

perhitungan untuk mengoptimalkan posisi *centroid* tersebut dengan cara melakukan perhitungan rata-rata untuk menghitung jarak antar *centroid* sampai terbentuk kluster baru dan *centroid* tersebut tidak lagi berubah posisinya [4].

Algoritma K-means tidak memiliki jaminan akurasi dikarenakan penentuan *centroid* awal dilakukan secara acak. Lalu berkembanglah algoritma ini menjadi *K-Means++* yang menjadi pembeda dengan algoritma sebelumnya, pembedanya adalah ketika memasukan data awal kluster data dimasukan pada setiap kluster dengan perhitungan jarak tertentu sehingga tidak lagi acak seperti K-Means yang sebelumnya [5]. Namun, setiap algoritma pasti memiliki kekurangan seperti halnya algoritma K-Means dan K-Means++. Hasil dari algoritma ini tidak dapat diinterpretasikan atau bukan merupakan data aktual dikarenakan data tersebut mengambil nilai rata-rata [6].

Algoritma K-medoids atau *partitioning around medoids* adalah algoritma yang dikembangkan dari algoritma K-Means sehingga secara keseluruhan kinerjanya masih sama dengan algoritma K-Means yaitu memecah *dataset* menjadi kelompok-kelompok dan keduanya sama-sama berusaha meminimalkan kesalahan. Cara kerja algoritma K-Medoids adalah pada saat menentukan posisi *medoid* menggunakan median sebagai media perhitungannya, sehingga hasil data tersebut merupakan data aktual dikarenakan mencari nilai tengah. Itulah yang menjadi pembeda antara algoritma K-medoids dengan algoritma sebelumnya [7].

Algoritma K-medoids telah diterapkan kedalam berbagai permasalahan penelitian terkait *clustering*, diantaranya adalah penelitian yang berjudul Implementasi Algoritma K-medoids Untuk *Clustering* Wilayah Terinfeksi Kasus Covid-19 di DKI Jakarta [8]. Penelitian tersebut dapat membantu pemerintah dalam menangani daerah yang memiliki kasus yang tinggi. Selain itu, ada juga penelitian dari STMIK Royal mengenai ekspor buah-buahan menurut negara tujuan menggunakan K-medoids sebagai algoritma *clustering* [9].

Pada penelitian ini, permasalahan yang diangkat dan dijadikan objek penelitian adalah ayat-ayat *Al-Qur'an* terjemahan bahasa Indonesia yang kemudian di klusterisasi menggunakan algoritma K-medoids. Penulis memilih *Al-Qur'an* sebagai objek penelitian dikarenakan *Al-Qur'an* memiliki keunikan yang mampu diklasifikasikan kedalam berbagai jenis seperti dikelompokkan berdasarkan kemiripan kata atau karakteristiknya.

Al-Qur'an merupakan kitab suci umat muslim yang sudah muncul dari berabad-abad lalu dan menjadi pedoman umat islam sekaligus menjadi sumber ajaran islam. Seperti halnya *Taurat* dan *Injil*, kata *Al-Qur'an* sendiri merupakan nama khusus untuk firman Allah SWT kepada para nabi-Nya, *Al-Qur'an* sendiri diturunkan untuk umat muslim yang kemudian diwahyukan kepada Nabi Muhammad SAW sebagai kitab suci terakhir. Secara etimologis, *Al-Qur'an* adalah *mashdar (infinitif)* dari *qara-a---yaqra-u—qira-atan-qur'a-nan_* yang berarti bacaan [10].

Hal ini berdasar pada firman Allah SWT yaitu :

إِنَّ عَلَيْنَا جَمْعَهُ وَقُرْآنَهُ (١٧) فَإِذَا قَرَأْنَاهُ فَاتَّبِعْ قُرْآنَهُ (١٨)

Artinya :

“*Sesungguhnya atas tanggungan Kami-lah mengumpulkannya (di dadamu) dan (membuatmu pandai) membacanya. Apabila Kami telah selesai membacakannya, maka ikutilah bacaannya itu.*” (Q.S Al-Qiyamah 75:17-18)

Al-Qur'an memiliki keistimewaan tersendiri sebagai kitab yang menjadi pedoman bagi umat muslim. Sebagaimana kitab-kitab sebelumnya, keistimewaan *Al-Qur'an* tersebut menjadi esensi *Al-Qur'an* sebagai kitab suci. Esensi itu adalah *hudan* yang berarti petunjuk untuk umat manusia khususnya seluruh umat muslim. Dengan *hudan* inilah seluruh anak cucu Adam dijamin hidup senang nan tenang di muka bumi [11].

Al-Qur'an dibagi ke dalam banyak ayat-ayat dan surat-surat. Para ulama menyepakati penempatan ayat dalam *Al-Qur'an* bersifat *tauqifi* yang didasarkan pada Riwayat *masyhur*, yakni dikutip oleh Quthb bahwa ketika ayat *Al-Qur'an*

diturunkan, Nabi menghimbau para sahabat dan berkata “Letakkan ayat ini pada surat yang menyebutkan begini dan begini” [12].

Al-Qur'an dapat diklasifikasikan menjadi banyak jenis, dikutip dari Syekh Muhammad ‘Abdul Azhim adz-Dzarqani oleh Dr.Wajidi Sayadi dalam artikelnya dikatakan bahwa *Al-Qur'an* dapat diklasifikasikan atas empat macam [13] :

1. Surat *Ath-thiwal*, yakni tujuh surat yang dikategorikan memiliki jumlah ayat yang panjang/banyak. Diantaranya yaitu *Al-Baqarah*, *Al-Imran*, *An-nisa*, *Al-Maidah*, *Al-An'am*, *Al'araf*, dan *Al-Anfal* digabung dengan *At-Taubah* karena tidak adanya perantara basmalah dikeduanya.
2. Surat *Al-Mi'un*, yaitu surat yang memiliki jumlah ayatnya 100 atau lebih sedikit.
3. Surat *Al-Matsani*, yaitu surat yang memiliki jumlah ayat kurang dari 100.
4. Surat *Al-Mufassal*, yaitu surat yang banyak fashl atau pemisahan antar surat dengan lafal Basmalah. Selanjutnya, Surat *Al-Mufassal* ini dikategorikan lagi berdasarkan panjang dan pendeknya surat.

Klasifikasi ayat-ayat *Al-Qur'an* tersebut mengartikan bahwa ayat-ayat dalam *Al-Qur'an* memang memiliki keterkaitan satu sama lain, baik itu dalam segi jumlah ayat bahkan pemisah diantara keduanya. Keterkaitan ayat-ayat *Al-Qur'an* tersebut juga membuktikan bahwa *Al-Qur'an* dengan segala keistimewaannya dapat dikelompokkan menjadi banyak jenis, bahkan salah satunya adalah dikelompokkan berdasarkan kemiripan katanya. Seperti contoh kata azab, surga, siksa, dan syaiton yang terdapat dalam ayat *Al-Qur'an* dapat dikelompokkan menjadi kelompok kata dengan karakteristik akhirat [14]. Selain itu, keterkaitan antara ayat satu dengan ayat yang lainnya dapat menjadi pembelajaran bagi umat muslim bahwasanya *Al-Qur'an* dengan segala keistimewaannya dapat dikelompokkan bukan hanya dari surat dan juz namun dapat juga dikelompokkan berdasarkan kemiripan katanya.

Penelitian sebelumnya mengenai *clustering* terhadap ayat-ayat *Al-Qur'an* terjemahan bahasa Indonesia telah dilakukan [14] [15] dengan kualitas *clusternya* adalah 0,274 untuk yang menggunakan algoritma k-means++ dan 0,3744 untuk yang menggunakan algoritma k-means. Pada penelitian kali ini, penulis menggunakan algoritma K-Medoids sebagai algoritma yang dipilih untuk proses *clustering* terhadap ayat-ayat *Al-Qur'an* terjemahan bahasa Indonesia.

Adapun perubahan yang dilakukan adalah ditambahkan langkah setelah pembobotan TF-IDF yaitu tahapan reduksi dimensi. Reduksi dimensi sederhananya adalah tahapan yang dilakukan untuk mengurangi dimensi atau kolom data sehingga data yang dimiliki memiliki ukuran yang relatif lebih kecil namun masih memiliki kualitas yang baik. Tujuan dari reduksi dimensi adalah untuk meningkatkan kualitas data sehingga akurasi yang dihasilkan nantinya bisa lebih baik [16].

Pengujian hasil kluster sebagai tahap evaluasi menggunakan metode *Silhouette Coefficient* yaitu sebuah metode evaluasi hasil kluster yang menghitung jarak setiap objek dalam sebuah kluster dengan kluster terdekatnya. *Output* dari nilai *Silhouette Coefficient* ini merupakan nilai *silhouette* dari keseluruhan data kluster sehingga nilai tersebut menjadi indikator untuk menilai kualitas hasil kluster [17].

Dengan demikian, penelitian ini akan melakukan *clustering* menggunakan algoritma K-medoids terhadap ayat-ayat *Al-Qur'an* terjemahan bahasa Indonesia sebagai objek penelitian yang diambil. Adapun, tahapan reduksi dimensi akan ditambahkan dalam proses *text processing* untuk mendapatkan nilai kualitas *clustering* yang lebih baik. Berdasarkan uraian tersebut, dirumuskanlah penelitian tugas akhir yang berjudul “**Kinerja Algoritma K-Medoids Untuk Clustering Ayat-Ayat Al-Qur'an Terjemahan Bahasa Indonesia**”.

1.2 Rumusan Masalah

Berdasarkan uraian latar belakang, didapat rumusan masalahnya yaitu sebagai berikut:

1. Bagaimana meningkatkan klusterisasi dokumen *Al-Qur'an* terjemahan bahasa Indonesia menggunakan algoritma K-medoids?
2. Bagaimana kualitas hasil *clustering* ayat-ayat *Al-Qur'an* terjemahan bahasa Indonesia menggunakan algoritma K-medoids?

1.3 Tujuan Penelitian

Selanjutnya, tujuan dari penelitian ini adalah :

1. Mengetahui cara meningkatkan klusterisasi pada dokumen *Al-Qur'an* terjemahan bahasa Indonesia menggunakan algoritma K-medoids sehingga dapat menjadi referensi untuk penelitian selanjutnya dalam melakukan klusterisasi ayat-ayat *Al-Qur'an* terjemahan bahasa Indonesia.
2. Mengetahui kualitas hasil *clustering* ayat-ayat *Al-Qur'an* terjemahan bahasa Indonesia menggunakan algoritma K-medoids.

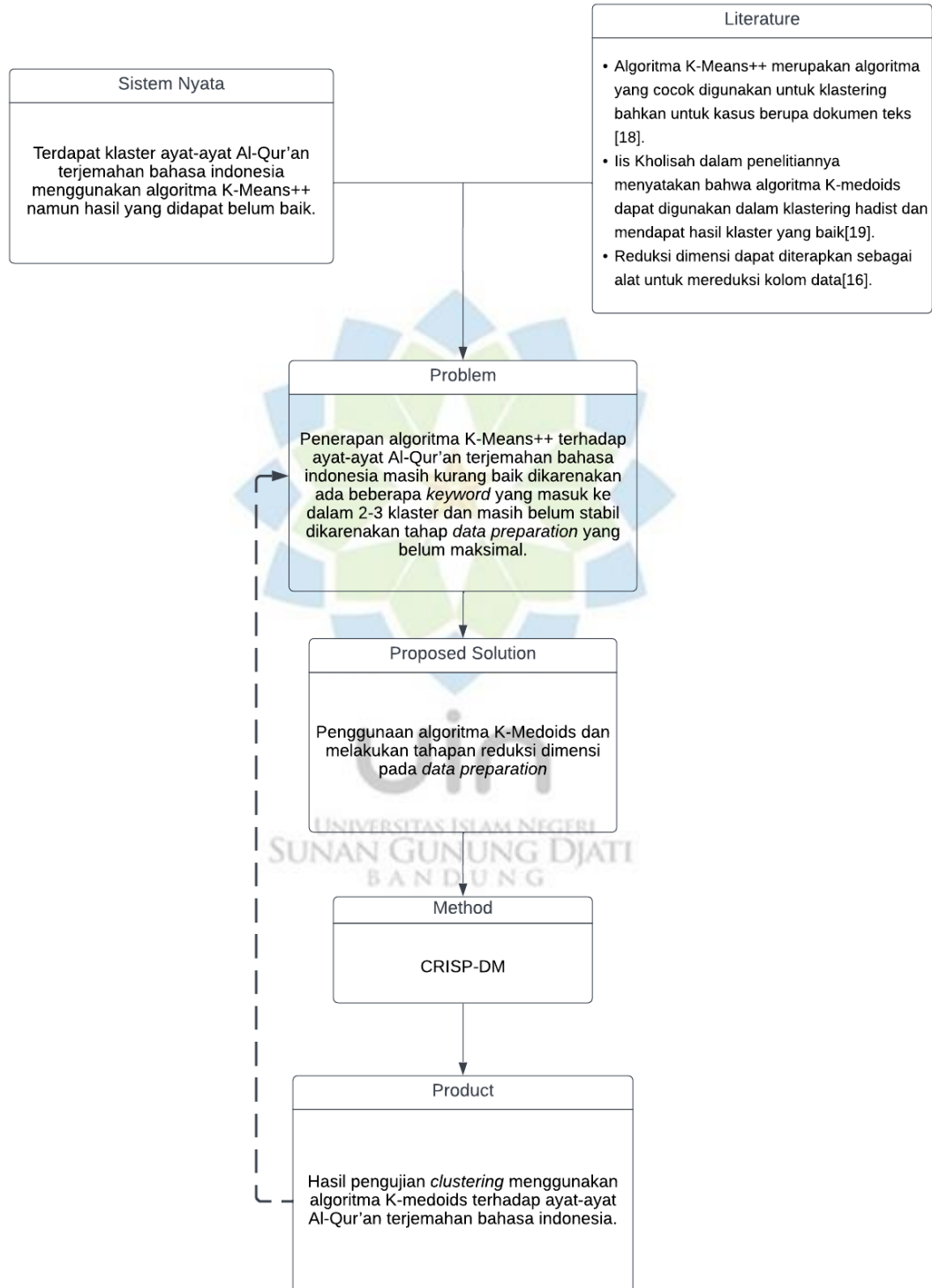
1.4 Batasan Masalah

Berikut adalah batasan masalah dari penelitian kinerja algoritma K-Medoids untuk *clustering* ayat-ayat *al-qur'an* terjemahan bahasa Indonesia:

1. Dataset yang digunakan adalah terjemahan berbahasa Indonesia bersumber dari website *quranenc.com* dan telah diterbitkan oleh Kementerian Agama Republik Indonesia. Dataset terjemahan ini sudah dikoreksi oleh pusat Terjemah Ruwwad [18].
2. Data yang akan di kluster berjumlah 114 surat atau 6236 ayat beserta terjemahan bahasa Indonesia.
3. Indikator kualitas hasil *clustering* menggunakan nilai *Silhouette Coefficient*.
4. Metode yang digunakan adalah Algoritma K-Medoids *Clustering*.
5. Tahapan penelitian tidak mencapai tahap deployment.

1.5 Kerangka Pemikiran

Kerangka pemikiran dari tugas akhir dapat dilihat dari gambar sebagai berikut :



Gambar 1. 1 Kerangka Pemikiran

Pada Gambar 1.1 mengvisualisasikan kerangka pemikiran yang menjadi landasan dilakukannya penelitian. Pada sistem nyata menyatakan adanya ketidaksesuaian antara penelitian yang telah dilaksanakan sebelumnya dengan literatur awal yaitu mengenai algoritma K-Means adalah algoritma yang cocok untuk klastering [19] lalu berdasarkan penelitian sebelumnya oleh Iis kholisah yang menggunakan algoritma K-Medoids dapat digunakan dalam klastering untuk topik hadist [20] selanjutnya pada proses klastering melakukan tahap reduksi dimensi sebagai alat untuk mereduksi kolom data [16]. Hal tersebut menyebabkan perlu adanya pembaharuan terhadap topik yang serupa namun dengan algoritma yang berbeda. Algoritma yang digunakan adalah K-medoids serta menambahkan tahapan reduksi dimensi untuk berupaya mengurangi waktu *modelling*. Dengan metode pengembangan CRISP-DM produk yang diinginkan adalah hasil pengujian *clustering* menggunakan algoritma K-medoids terhadap ayat-ayat *Al-Qur'an* terjemahan bahasa Indonesia.

1.6 Sistematika Penulisan

Sistematika penulisan yang disusun pada penelitian yang berjudul Kinerja Algoritma K-Medoids Untuk *Clustering* Ayat-Ayat *Al-Qur'an* Terjemahan Bahasa Indonesia ini terdiri dari 5 bab, masing-masing bab yang dirancang memiliki tujuan tersendiri yang ingin dicapai. Berikut adalah sistematika penulisan yang dirancang :

BAB I PENDAHULUAN

Pada bab pertama berisi latar belakang penelitian, rumusan masalah penelitian, tujuan penelitian, batasan masalah penelitian, kerangka pemikiran penelitian, sistematika penulisan.

BAB II KAJIAN LITERATUR

Bab kedua berisi pembahasan mengenai teori, konsep, model dan rumus yang sedang dikaji serta memiliki hubungan untuk mendukung implementasi dari penelitian yang dilakukan.

BAB III METODOLOGI PENELITIAN

Pada bab ketiga merupakan pembahasan tentang metodologi penelitian yang digunakan yaitu CRISP-DM. Pembahasan tersebut meliputi *business*

understanding, data understanding, data preparation, modelling, dan evaluation.

BAB IV HASIL DAN PEMBAHASAN

Bab keempat adalah pembahasan dan pemaparan mengenai hasil klusterisasi menggunakan algoritma yang digunakan serta evaluasi dari hasil klusterisasi tersebut.

BAB V PENUTUP

Setelah melaksanakan penelitian, maka disampaikan kesimpulan dan juga saran yang diajukan untuk peningkatak penelitian ini sebagai bagian dari bab terakhir yaitu penutup.

